Scalability Analysis of Tightly-coupled FPGA-cluster for Lattice Boltzmann Computation

Yoshiaki Kono, Kentaro Sano, Satoru Yamamoto
Graduate School of Information Sciences, Tohoku University, Japan

Outline

- Introduction
- Lattice Boltzmann method (LBM)
- FPGA-cluster for LBM
- Performance model & analysis
- Conclusions
**Introduction**

- **Lattice Boltzmann method (LBM)**
  - method to compute fluid dynamics
  - High parallelism
  - Low operational intensity (each op. requires many data)

For large-scale parallel computing....

- **Today’s micro-processors**
  - cannot provide peak-performance.
  - memory bandwidth is insufficient.
  - Limited scalability in large scale sys., conspicuous for strong scaling.
  - caused by imbalanced performance and bandwidth.

- **Custom-computing machine**
  - High utilization if we design appropriate HW
    (performance & bandwidth are balanced)

---

**FPGA-based Custom Computing for HPC**

- FPGAs have been getting larger and faster
  - High-speed I/O (Tx,Rx) ex. LDVS
  - Higher bandwidth memories and chip-I/Os

Promising devices for custom HPC
Related Work: Custom LBM Machines

[Sano2007] FPGA-based streaming computation for LBM

[Murtaza2009] Custom-computing machine for cellular-automata

[Murtaza2011] LBM computation on Maxwell (multi-FPGA system)

Tightly-coupled FPGA cluster with a dedicated network

Objectives of this Research

Strong-scalability analysis of tightly-coupled FPGA-cluster

- Direct connection between FPGAs beyond the node

Architecture design for scalable LBM on FPGA-cluster

- Spatial- and Temporal- Parallelism

Contributions of this paper

- Scalable-architecture design for LBM computation
- Sustained-performance model
- Analysis of strong-scalability
Lattice Boltzmann method: LBM

Initialize

collision of particles

propagation of particles

N

Y

finish?

particle densities for 9 directions

propagation of particles (gather)

movement of densities

Processing Element for Stream Computation

traverse

C

collision of particles

particle densities for 9 directions

propagation of particles

movement of densities

PE

lattice-points

Collision Unit

Delay Buffer

Propagation Unit

updated points
Processing Element for Stream Computation

traverse

PE

Collision Unit

Delay Buffer

Propagation Unit

updated points

lattice-points

collision of particles

particle densities for 9 directions

propagation of particles

movement of densities

particle densities for 9 directions

movement of densities
Processing Element for Stream Computation

Spatial parallelism
✓ sub-lattices can be computed in parallel

Temporal parallelism
✓ stream-computations can be pipelined along time-steps
PE Array for Spatial and Temporal Parallelisms

Spatial parallelism

Temporal parallelism

Large PE-Array on Tightly-Coupled FPGAs

- 1D-ring network for higher bandwidth between FPGAs
- 1D decomposition of a large PE-array

Tightly-coupled FPGA-cluster

Accelerator-domain network (ADN)
Prototype FPGA-Cluster

Peck: 6.4GB/s

FPGA-1
FPGA-2
FPGA-3
FPGA-4
FPGA-5

DDR2
DDR2
DDR2
DDR2
DDR2

Host 1
Host 2

PCI Express

Parameters of FPGA-Cluster

\[ W_{\text{mem}} \leq \frac{N_{\text{PE}}}{2} \times B \] [bytes]

\[ W_{\text{mem}} \] : Memory bandwidth
\[ W_{\text{I/O}} \] : Network bandwidth
\[ B \] : BRAM size that can be used as buffers for propagation
\[ N_{\text{PE}}^{\text{max}} \] : Max num. of PEs that can be implemented on each FPGA

These Parameters influence Performance and Scalability
Performance Limited by Resources (# of PEs)

- Assuming 100% utilization,
  \[(\text{Peak performance of each FPGA}) \propto N_{PE}^{max}\]
- For number of FPGAs \(N_{\text{FPGA}}\),
  \[(\text{Total Performance}) = O(N_{PE}^{max} \times N_{\text{FPGA}})\]

Performance Limited by Network-Bandwidth

- To utilize PEs for 100% cycles, communication time must be hidden with computing time.

\[(\text{comp. time}) = \frac{(\text{lattice size})}{(\text{performance})} \leq (\text{comm. time}) \left( \propto \frac{1}{W_{IO}} \right)\]
\[(\text{performance}) \leq \alpha(W_{IO})\]
No Limitation by Memory-Bandwidth

Higher memory-bandwidth

- More PEs for spatial parallelism

Lower memory-bandwidth

- More PEs for temporal parallelism

Total performance is not limited by memory-bandwidth

Performance Limited by Buffer Size

- When using fewer FPGAs, a larger sub-lattice is allocated to each PE in strong-scaling.
- We must reduce \( N_{PE} \) so that each PE has a larger buffer.
- Less FPGAs, and less \( N_{PE} \) on FPGA

\[
\text{Performance} \propto \# \text{ of PEs} = O\left( (\text{Number of FPGAs})^2 \right)
\]
Parameters Used for Analysis

We designed LBM PEs for single-precision by using Flopoco [dinechin 2011].

<table>
<thead>
<tr>
<th>Design / FPGAs</th>
<th>$N_{\text{PE}}^{\text{max}}$</th>
<th>$B$ [MB]</th>
<th>$W_{\text{mem}}$ [GB/s]</th>
<th>$W_{\text{I/O}}$ [GB/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratix IV EP4SGX230 (DE4)</td>
<td>6</td>
<td>1.6</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Stratix V 5SGXB6</td>
<td>6</td>
<td>6.6</td>
<td>24</td>
<td>116</td>
</tr>
<tr>
<td>Stratix V 5SGSD8</td>
<td>26</td>
<td>6.4</td>
<td>36</td>
<td>84</td>
</tr>
</tbody>
</table>

For Stratix IV

✓ estimate memory- and network- bandwidth available on DE4 Board

For Stratix V

✓ estimate parameters with a spec sheet

Strong-Scalability Analysis (4000 × 4000 Lattice)

- 100% condition at 1-FPGA
- Stratix V have 4 times higher performance than Stratix IV
Strong-Scalability Analysis (4000 × 4000 Lattice)

Network bandwidth is most important for strong-scalability of FPGA-cluster. For FPGA cluster with enough bandwidth, we achieve complete strong-scalability up to thousands of FPGAs.
Conclusions

Strong-scalability analysis of tightly-coupled FPGA-cluster
Architecture design for scalable LBM on FPGA-cluster

- Stream architecture for LBM computation
  - Exploit spatial- and temporal-parallelism of stream-computation
  - Balance arithmetic-performance with bandwidth
- Strong-scalability analysis for the tightly-coupled FPGA-cluster
  - Formulate performance model and estimate performance of the cluster
  - Single Stratix V has 445GF
    (GeForce GTX-590 has peak performance of 2.488 TFLOPs)
  - Stratix V can scale up to 666 FPGAs, delivering 296TF

We can achieve good scalability if FPGA cluster has ADN with sufficient bandwidth.

Future Work

- Implementation and benchmarking
- Performance model for 3D LBM computation
- FPGA Cluster with multi nodes of Stratix V FPGAs